MTR 070281

MITRE TECHNICAL REPORT

Quality Assessment of Trustworthiness of AFMC Acquisition Data

September 2007

Herman L. Karhoff - Investigator

Sponsor:	554th Electronic Systems Wing	**Contract No.:**	FA8721-07-C-0001
Dept. No.:	E141	**Project No.:**	0307754001

MITRE
Corporate Headquarters
McLean, Virginia

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
SEP 2007		**00-00-2007 to 00-00-2007**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Quality Assessment of Trustworthiness of AFMC Acquisition Data	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **MITRE Corporation,7515 Colshire Drive,McLean,VA,22102-7539**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

To become a more efficient and effective joint-expeditionary force, the Air Force (AF) and its business partners are adopting an enterprise view that optimizes resources (i.e., people, process, and technology). To achieve this end, the AF has embarked upon an aggressive enterprise Information Technology (IT) modernization strategy. A major challenge with the planning and implementation of transformation/migration strategies is the ability to determine the quality of AF data. A basic premise is that data of unknown quality is inherently untrustworthy. Few would disagree with the premise that good quality data (i.e., timeliness and accuracy) is critical to aiding AF leadership in making the right decisions. How can these decision makers trust the data if they do not have a means to assess and measure their data quality? Is it impossible to measure something that is not understood, and how do you manage something that cannot be measured? These are some of the key questions MITRE seeks to answer in this Mission Oriented Investigation and Experimentation (MOIE) initiative. The purpose of this paper is two-fold: (1) to heighten awareness on the importance and impacts of data quality (DQ); and (2) to document (i.e., DQ assessment methodology, measurement techniques and assessment criteria) the findings and outcomes from this MOIE research. The approach is to apply semantics and heuristics (i.e., utilization of architectures, methodologies, state-of-the art software tools, implementation techniques, and production test data) in exploring capabilities that enhance data quality and thus improve the quality of decisions made with enterprise data. AF production data (e.g., invoice transactions) was used to verify and validate the MOIE hypothesis, assumptions, and findings.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**39**	

Abstract

To become a more efficient and effective joint-expeditionary force, the Air Force (AF) and its business partners are adopting an enterprise view that optimizes resources (i.e., people, process, and technology). To achieve this end, the AF has embarked upon an aggressive enterprise Information Technology (IT) modernization strategy. A major challenge with the planning and implementation of transformation/migration strategies is the ability to determine the quality of AF data. A basic premise is that data of unknown quality is inherently untrustworthy. Few would disagree with the premise that good quality data (i.e., timeliness and accuracy) is critical to aiding AF leadership in making the right decisions. How can these decision makers trust the data if they do not have a means to assess and measure their data quality? Is it impossible to measure something that is not understood, and how do you manage something that cannot be measured?

These are some of the key questions MITRE seeks to answer in this Mission Oriented Investigation and Experimentation (MOIE) initiative. The purpose of this paper is two-fold: (1) to heighten awareness on the importance and impacts of data quality (DQ); and (2) to document (i.e., DQ assessment methodology, measurement techniques and assessment criteria) the findings and outcomes from this MOIE research. The approach is to apply semantics and heuristics (i.e., utilization of architectures, methodologies, state-of-the art software tools, implementation techniques, and production test data) in exploring capabilities that enhance data quality and thus improve the quality of decisions made with enterprise data. AF production data (e.g., invoice transactions) was used to verify and validate the MOIE hypothesis, assumptions, and findings.

Table of Contents

List of Figures

1 Background

The Air Force (AF) and other Department of Defense (DoD) organizations (e.g., Defense Logistics Agency) are spending billons of dollars to replace legacy systems with Enterprise Resource Planning (ERP) solutions. Current evidence indicates that poor Data Quality (DQ) is pervasive and is having a significant impact on AF operations[1]. This is largely attributed to DQ issues. It is a proven fact that ERP solutions require highly accurate data (e.g., bills of material (90%), routings [production lead times] (98%), inventory (90%), purchase order status (95%). Lessons learned from the Navy and Army (i.e., the Army determined the accuracy of their legacy system data to be 68%[2]) noted data quality can be a major obstacle to ERP deployment.

This begs a number of questions, e.g., what is the value and fidelity of the current legacy system data? What are the impacts of DQ issues (e.g., migration into an ERP solution, cost to fix, and management burden)? How can we measure or quantify the answers to these questions? What course of action should the AF pursue to ensure the right material and services are at the right place at the right time? The answer to these questions can only be understood once the Data Quality Requirements (DQR) are known; once a DQ assessment of legacy system data is completed, and the essence or root causes of the DQ problems are been identified.

One of the intents of Section 515 of the Treasury and General Government Appropriations Act for Fiscal Year 2001 (Public Law 106-554; H.R. 5658) was to address this issue. This Act directed the Office of Management and Budget (OMB) to issue government-wide guidelines to "provide policy and procedural guidance to Federal agencies to ensure and maximize the quality, objectivity, utility, and integrity of all information (including statistical information) disseminated by Federal agencies". By October 1, 2002, agencies were required to issue implementing guidelines to ensure compliance. OMB's emphasis on data being the critical asset of the information age has also been acknowledged in the commercial sector. Lou Gerstner, Executive Officer at IBM, illustrates the point, "Inside IBM, we talk about ten times more connected people, 100 times more network speed, 1,000 times more devices and a million times more data"[3] The Deputy Secretary of Defense, in a memorandum[4] to the Joint Chiefs and Department Heads, noted, "DoD Components shall adopt standards of quality that are appropriate to the nature and timeliness of the information they disseminate". Subsequent DoD Guidelines on Data Quality Management (DQM)[5] have been issued that define the high-level processes (See Figure 1) as they relate to the five steps of Six Sigma– yet the DQ problem persists.

---

[1] *DLA/SBSS Backorder Data Validation Study*, Fred Nannarone, Dynamics Research Corp, 2005

[2] Data Quality Best Practices & Lessons Learned, by Michael Gallagher, Site Readiness Manager, Army LMP Program, 31 July 2007

[3] McDougall, Paul. "More Work Ahead," Information Week, December 18-25, 2000, p.22.

[4] Memorandum dated February 10, 2003, Subject: *Ensuring Quality of Information Disseminated to the Public by the Department of Defense*

[5] *DoD Guidelines on Data Quality Management*, http://ssed1.ncr.disa.mil/srp/datadmn/dqpaper.html.

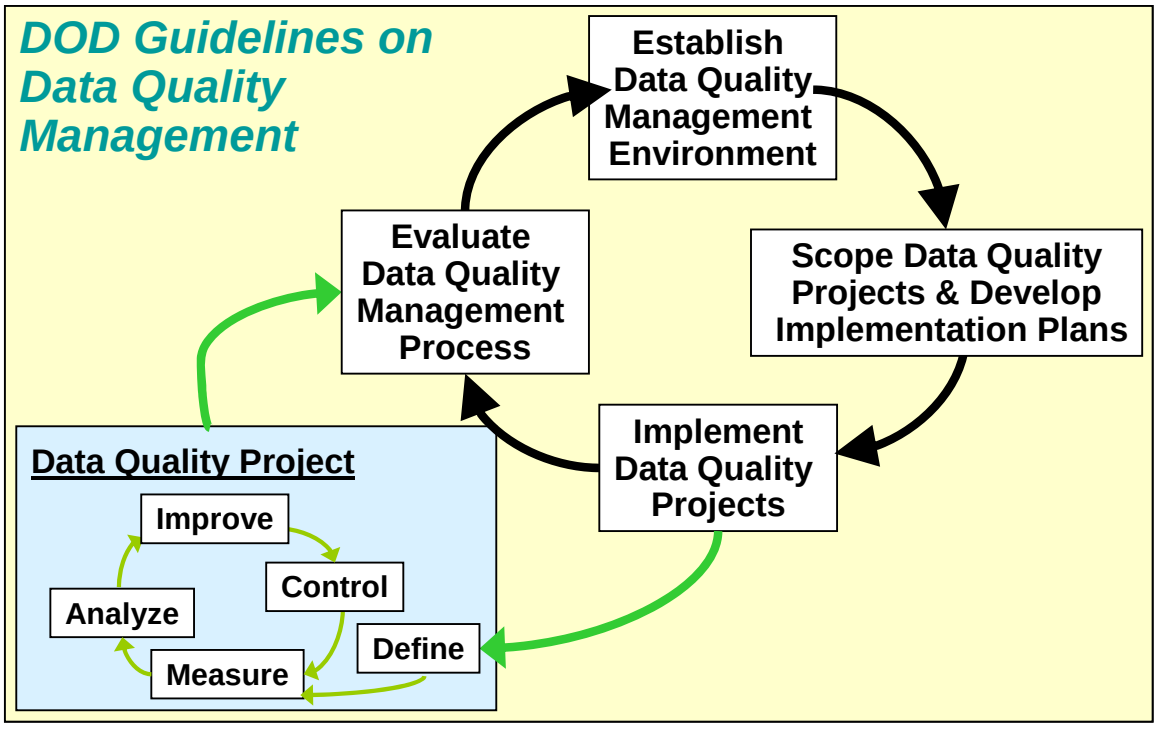


Figure 1. DoD Data Quality Management

Compounding the DQ challenges is the loss of intellectual capital, because of personnel reductions and retirements. Experts with the knowledge and understanding of the business processes (i.e., requirements and business rules), data of the legacy systems are dwindling. Replacement resources lack the training and/or the desire to continue working with legacy system mindsets. They become overwhelmed with not knowing the cause and sources of the DQ problem. When they do attempt to work the issues, it is usually limited in focus to their business domain or to their local needs, not from an AF or DoD enterprise perspective. Plus, there is a large amount of AF data which has its origin from external sources (e.g., contractors and other Service Agencies), entities for which the AF has little or no influence over DQ practices.

One might ask, why is it so hard, and how does the AF go about addressing these mandates in a time when resources are limited due to the increase in global operations tempo? What are the costs and inefficiencies of accepting DQ issues as a fact of life? Are there additional costs that lead to fabricating work-around procedures or processes? How is this adding to the burden of AF maintenance production, supply chain management, and acquisition, and payment of goods and services?

Tool vendors will suggest their software solutions provide the answers to many of the above questions. They offer a variety of DQ tools, but which tools are best suited to help work these challenges? For example, Enterprise Application Integration (EAI) and Extract and Load (ETL) tools offer help in the areas of data profiling, cleansing, and translation. Unfortunately, these tools are tailored for a specific purpose, have proprietary restrictions, and frequently have long learning curves. These tools may help facilitate the DQ analysis, but they will not clean up the data.

2 Researching Solutions

In this Mission Oriented Investigation and Experimentation (MOIE) initiative, MITRE proposes to apply semantics and heuristics to the expression of DQ to improve the trustworthiness of enterprise data. MITRE's research is based on the following premises:

- Data of unknown quality is inherently untrustworthy
- Conversely, data of known quality can be treated appropriately by DQ management tools
- The AF depends on data that is manipulated and aggregated by hundreds of legacy systems
- Air Force data lacks ontology and semantics needed for trustworthy data processing and decision making.

2.1 Research and Assessment Objectives

The objective of this MOIE initiative is to architect, capture, test, and apply the semantics of DQ, using various tools in a net-centric, distributed Service Oriented Architecture (SOA) environment.

It is important to note that sources of information are frequently diverse, disparate, and often collected from heterogeneous mix of data sources. These conditions may be tolerated at the operational level due to the differing need for data to support local decision making. However, inconsistencies in the interpretation, business rule application, etc., attribute to semantic DQ issues. As a result, the true meaning of the data provided to enterprise decision-makers can be hard to ascertain. We propose to clarify the meaning of the data by creating semantic constructs for DQ such as accuracy, consistency, completeness, timeliness, pedigree, etc. Furthermore, by creating heuristics that use these values, we can now start to assess and measure DQ. The data may then be summarized and presented to the end-user in a way that allows them to interact at the decision level rather than at the data level.

2.2 Methodology

Extensive research by David Marco and Michael Jennings[6] suggested there is an underlying metadata model for metadata management. However, how does one apply semantics and heuristics techniques in the expression of DQ to improve the quality of decisions made with enterprise data? In pursuit of answers to this and other questions noted in Section 1, the MOIE Team concluded the following steps were necessary for the verification and validation of existing theories, concepts, and comparative analysis of the considered approaches:

- Define DQ
- Create DQ architecture
- Create metadata models (i.e., conceptual, logical, and physical) that represent the semantics of DQ (i.e., identify the inherent characteristics and attributes of DQ)
- Capture AF legacy system data (i.e., invoice transactions) as a baseline (i.e., notional use case scenarios) to test, verify and validate MOIE concepts
- Identify DQ reporting requirements and create dashboard DQ reports
- Design and build a DQ infrastructure foundation utilizing emerging web tools (e.g., XML, RDF, and OWL) executing in a distributed SOA environment
- Test and demonstrate a systematic approach to representing, measuring, capturing and using DQ information using AF invoice data

2.3 Define Data Quality

One finding of interest from this study is the number of different perspectives and interpretations of DQ. OMB defines quality as an encompassing term comprising utility, objectivity, and integrity. These guidelines frequently refer to these statutory terms, collectively, as "quality".[7]

> **Utility -** refers to the usefulness of the information to its intended users, including the public. In assessing the usefulness of information that the agency disseminates to the public,
>
> **Objectivity -** consists of two distinct elements: presentation and substance. The presentation element includes whether disseminated information is presented in an accurate, clear, complete, and unbiased manner and in a proper context. The substance element involves a focus on ensuring accurate, reliable, and unbiased

[6] Publication: *Universal Meta Data Model*, Wiley Publishing, Inc, 2004

[7] Office of Management and Budget; *Guidelines for Ensuring and Maximizing the Quality, Objectivity, Utility, and Integrity of Information Disseminated by the Federal Government;* Republication, 67 Fed. Reg. 8452 (Feb. 22, 2002)

information. In a scientific, financial, or statistical context, the original and supporting data will be generated, and the analytic results will be developed, using sound statistical and research methods.

Integrity – refers to security, the protection of information from unauthorized access or revision, to ensure that the information is not compromised through corruption or falsification.

For the purposes of this MOIE, we define DQ in the context to the quality of data. Data is of high quality "if they are fit for their intended use in operations, decision making and planning".[8] Alternately, the data is deemed of high quality if correctly represented in a real-world construct or context in which the data is used. Figure 2 portrays some of the desired characteristics and features of "data that's fit for use". Dr. Redman recommends every organization have a DQ vision. It is one thing to have a vision, however achieving the vision may be illusive unless there is a commitment to ensure DQ objectives are well defined and executed.

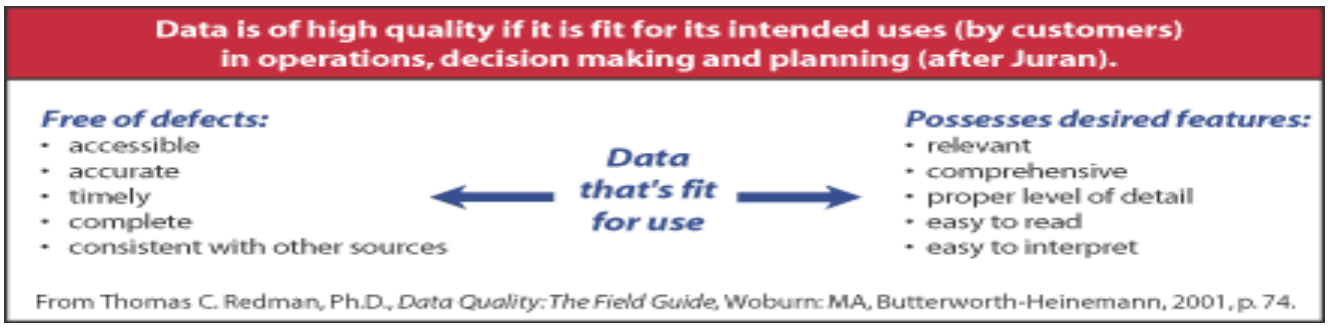


Figure 2. Data Quality

2.3.1 Creating a Data Quality Architecture

Before tackling any problem it is a good practice to architect the major components which comprise the proposed solution to the problem. The four major components of Data Quality Architecture (DQA) are: (1) the Information Manufacturing System, (2) Data Management Tools, (3) Metadata Repository, and (4) DQ Ontology/Metamodel (See Figure 3).

[8] Wikipedia; definition by J.M. Juran

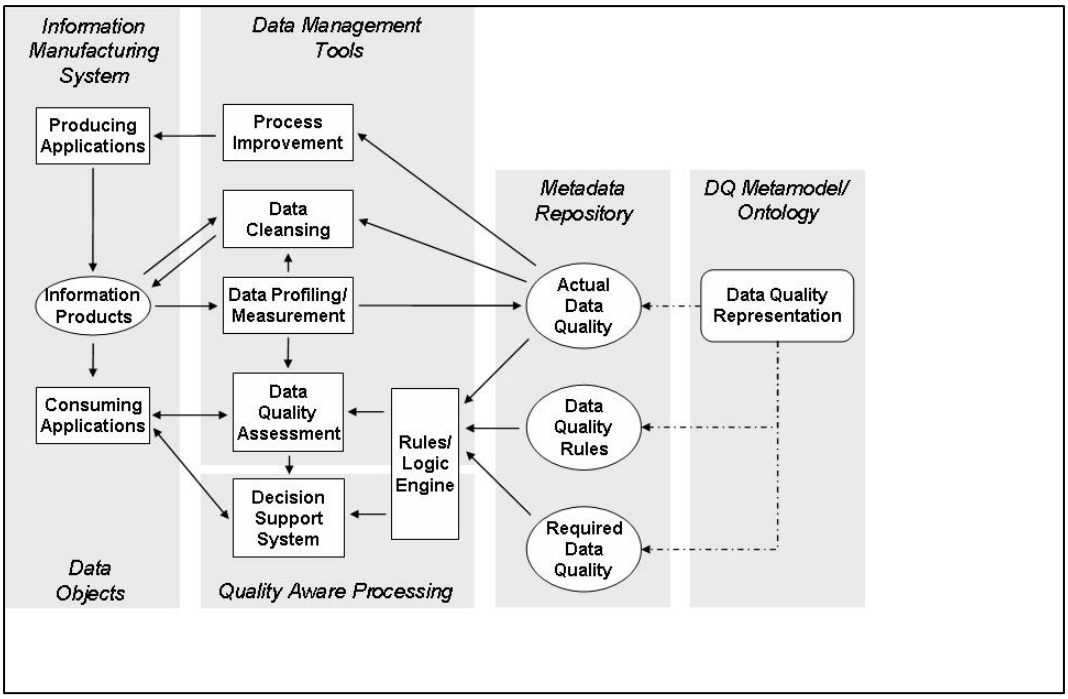


Figure 3. Data Quality Architecture

The Information Manufacturing System is representative of all of the AF business applications used to collect, assemble, produce, or manufacture (similar to manufacturing activities) the information required to execute to the AF's mission. For example, the AF develops demand and supply plans to forecast what resources (i.e., material, people, facilities, etc.) are required to maintain, repair, and overhaul an aircraft. The AF Centralized Spare Parts Forecasting System, D200A, based on numerous inputs (e.g., flying hours, material failure rates, repairable carcasses in the supply chain, and due-ins from suppliers), determines the material requirements in support of depot maintenance workload planning and scheduling. Assuming the projected item demand exceeds existing inventory, it may be necessary to initiate a Purchase Request to contract for item shortfalls (consuming applications). It is important to note that DQ issues can occur at any point in the information or data flows between producing and consuming applications. If the planning data is wrong; time, money, and production performance is lost. This impacts the supply chain (i.e., promotes the need for the inclusion of additional assets in the supply pipeline), production scheduler, the operation managers, decision makers, and eventually the warfighter. In addition, the DQ issues are compounded by frequent manual data manipulations and automated extrapolations by being recycled back into the next iteration of the information manufacturing lifecycle.

The data management tools component includes those capabilities (i.e., methodologies, tools, [e.g., ETL, EAI, and DQM], e-business integration broker, and enterprise application integration tools) used as aids for data profiling, DQ measurement and assessment, and data cleansing.

The metadata repository is a database that is populated with data descriptions, business rules, quality metrics, measurement and assessment criteria along with the actual DQ measurements, and assessments of the data objects of interest. The DQ ontology/metamodel is the underpinning metadata structure that organizes the content of the Metadata Repository.

2.3.2 Creating Data Quality Meta-models (Conceptual, Logical, and Physical)

Meta-modeling is the construction of a collection of concepts (e.g., objects, things, and terms). It is an explicit model of the constructs and rules needed to build a DQ metamodel. Metamodeling provides the ontology of the DQ domain by using entities in an entity-relationship-attribute or object-oriented modeling framework as the foundation for metadata (i.e., data about data) integration. A meta-model can be viewed from different perspectives such as:

- A set of building blocks and rules used to build models
- A model of a domain of interest
- An instance of another model

There is an assumption that every transaction processing action has a consequence, some are positive, others are negative. It was also assumed that organizations have a vested interest in understanding the impacts of change (i.e., business process re-engineering changes and application changes). The DQ metamodel will provide multiple pathways of insight into understanding the subtleties of these impacts by profiling (e.g., how systems process, interpret, manipulate, and transform) the data from the point of system entry and throughout the transaction processing lifecycle. This allows both business and technical users the ability to make informed decisions on what changes offer the best opportunity to improve DQ and the overall efficiency of the organization.

The first major effort was to develop a conceptual DQ metamodel (See Figure 4). Essential to building a conceptual data metamodel is to model the highest level of entities and their relationships. This model does not include any data attributes, primary or foreign keys, it is only a framework for lower level models.

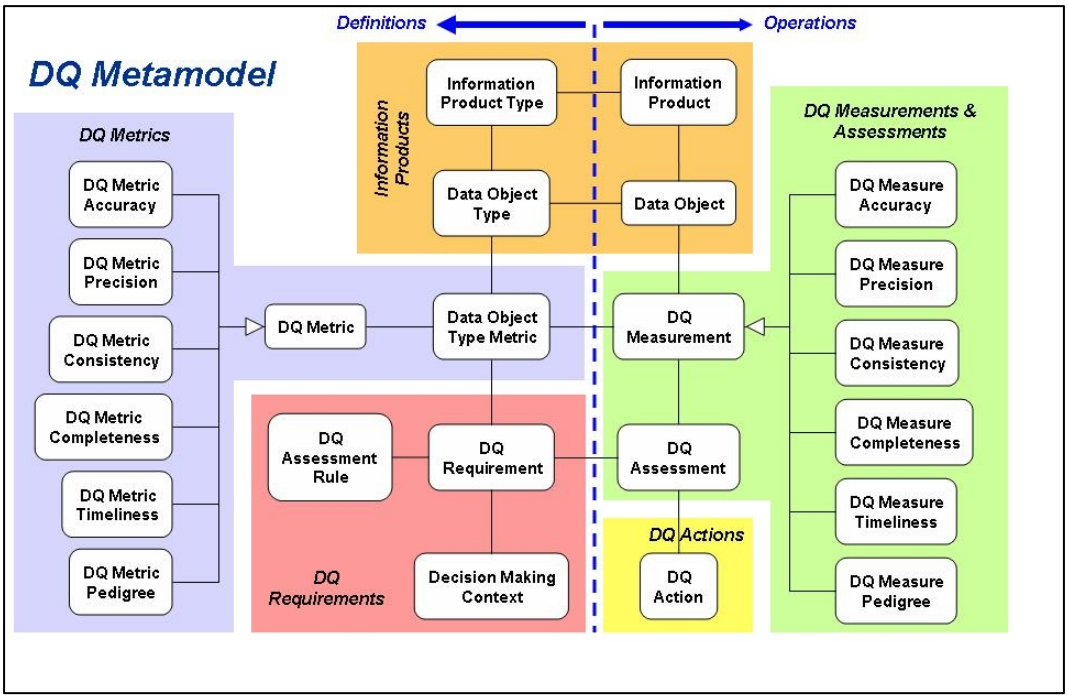


Figure 4. Conceptual Data Quality Meta Model

The next step was the creation of a logical DQ metamodel. Features of the logical DQ metamodel include all of the lower level entities and relationships, modeled in third normal form. All attributes for each of the entities are specified along with their primary and foreign keys. At this level, we described the data in as much detail as possible, without regard to physical implementation considerations (e.g., database performance and transaction handling).

The most challenging part of the DQ logical modeling effort was identifying the information about the data (i.e., contextual definitions, dimensions, and attributes) to be modeled. Extensive research by Wang and Strong has identified over 179 DQ attributes.[9] Due to the impractically of this number of attributes, they first proposed 20 DQ dimensions, which where assembled into four categories and eventually pared down to 15 dimensions[10]. Wang and Strong demonstrated that 15 dimensions were relevant to a generic population (generic in the sense that the study did not sample any specific domain or industry, but a large number of domains and industries). The purpose of referencing the Wang and Strong DQ framework is to inform the reader of the depth and breadth of what can be considered when performing a DQ assessment. Extending the research into the other dimensions noted in Figure 5 would have required surveys of an array of

[9] Beyond Accuracy: *What Data Quality Means to Data Consumers*, Journal of Management Information Systems, 1996.
[10] *Introduction to Information Quality*, by Fisher, Lauria, Chengalur-Smith, and Wang, 2006

users who are involved in the processing and handling of data which was beyond the scope and intent of this MOIE.

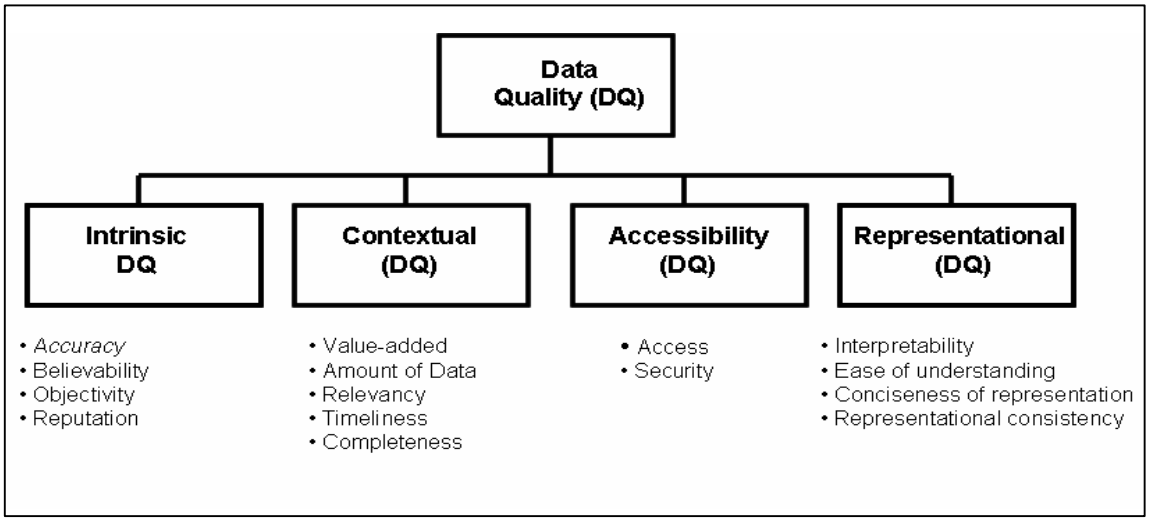


Figure 5. Data Quality Hierarchy of Four Categories and 15 Dimensions

Upon determining the required DQ attributes and taxonomies (e.g., DQR, DQ metrics, assessment and measurement criteria, decision making context, confidence factors), the resulting logical DQ metamodel was architected in Metastorm's ProVision modeling tool. The resulting architecture artifacts were used as a premise for metadata environment standardization and requirement specifications for configuring the system and technical architectures to be used to support this research. A separate whitepaper[11] provides a detailed interpretation of the logical DQ metamodel for this MOIE.

For this MOIE effort, we elected to focus on the DQ metrics/dimensions of accuracy, precision/uncertainty, timeliness, completeness/brevity, consistency, and pedigree/lineage. The following are logical model interpretations:

2.3.2.1 Accuracy

Definition*:* Degree to which the reported information value is in conformance with the true or accepted values.

DQ Requirement*:* For reporting and tracking purposes, each shipment shall be accompanied by a certificate of compliance (DD250 Invoice Document from which the invoice transaction has its origin) signed by a company and government official responsible for product assurance. The certification shall identify as a minimum the purchase order number, line item number, part

[11] Bill McMullen, *A Flexible and Generic Data Quality Metamodel*, August 2007

number as listed on the Purchase Order, quantity, and the manufacturer in accordance with (IAW) the Contract Terms and Conditions. Provide the ability to identify and count all invoice error conditions by processing location (i.e., Air Logistic Centers [ALC]).

Transaction Business Rule Violation*:* See business rule for invoice transactions in Appendix A.

Measurement Rule: Invoice error count by location = error count by location +1; for Error Type (n) not = blank, where n = 1 to 3.

Measurement: Is related to invoice error counts by processing location.

Assessment Rule: Percent invoice errors by processing location >= average percent errors from all processing locations.

Assessment: Red means >= Standard Deviation (SD) from average on high side; yellow means < SD from average, but > average; and green means =< average.

2.3.2.2 Precision/Uncertainty

Definition: Exactness or confidence in value (vs. imprecise, uncertain, approximate, probabilistic, or fuzzy).

DQ Requirement: Ability, from a system invoice processing perspective, to produce the same value or result, given the same input conditions and operating in the same environment.

Transaction Business Rule Violation: J041 invoice processing logic should be in accordance with DoD and AF governance.

Measurement: Percent of non-compliance occurrences.

Measurement Rule: Count of the occurrences as to when J041 fails to achieve the DQR.

Assessment Rule: Percent of invoices processing in J041 = non-compliance with governance.

Assessment: Red means high degree of uncertainty, green means 99.8% of invoices are in accordance with DoD and AF governance mandates.

2.3.2.3 Timeliness

Definition: The degree of freedom from variation or contradiction. The degree of satisfaction (i.e., compliance to DQ constraints [including format, syntax, and structure]) resulting from getting information to the right person and location in a timely manner.

DQ Requirement: All invoices are to be processed within a specific time period to allow payments without incurring interest penalties and accurate reporting of vendor performance.

Transaction Business Rule Violation: Vendor payment requirements are established by the Terms and Conditions of the Contract, usually within 30 days after receipt of goods and services. Contractor performance (i.e., J041 business rule) is determined by not having received a vendors invoice(s) with a 15-day period after the delivery schedule due date.

Measurement: Percentage breakout by age of invoice error by processing location.

Measurement Rule: Age (by location) = date-of-processing minus date-of-invoice transaction.

Assessment Rule: Average age of errors by processing location >= 5 days.

Assessment: Age of invoice error by processing location.

2.3.2.4 Completeness/Brevity

Definition: Degree to which values are present in the attributes that require them. Degree to which values not needed for decision making is excluded.

DQ Requirement: All invoice required fields must contain the information prescribed in the Terms and Conditions of the Contract (e.g., SHIP TO/Mark FOR).

Transaction Business Rule Violation: See business rules for invoice business rules.

Measurement: The count of incomplete invoices not in accordance with current governance. The percentage of items that are in violation of established business rules.

Measurement Rule: Number of incomplete invoices = sum invoices with null values in required fields.

Assessment Rule: Percent of incomplete invoice errors by Processing Location >= average percent errors from all processing locations.

Assessment: Red means >= SD from average on high side; yellow means < SD from average, but > average; and green means =< average.

2.3.2.5 Consistency

Definition: Degree to which specified data values are up to date.

DQ Requirement: Governance provides periodic changes in the domain range of values for invoice data (e.g., Unit of Issue, Mode of Shipment). Consistent application of this governance shall be applied to invoice creation and processing.

Transaction Business Rule Violation: See domain range of values (i.e., business rules) for invoice transactions in Appendix A.

Measurement: The count of incomplete invoices not in accordance with current governance. The percentage of items that are in violation of established business rules.

Measurement Rule: Number of invoice errors = invoice content not in accordance with current governance.

Assessment Rule: Percent of invoice errors by processing location >= average percent errors from all processing locations.

Assessment: Red means >= SD from average on high side; yellow means < SD from average, but > average; and green means =< average.

2.3.2.6 Pedigree/Lineage

Definition: Degree to which specified data values are up to date. The history of data origin (also called lineage or provenance), ancestral relationships, and subsequent transformation(s) over its lifecycle.

DQ Requirement: Properly identify and record the input date/time stamp, input source, and processing location of all invoices.

Transaction Business Rule Violation: All invoices must be assigned (J041 assigns) a Source Code, Date of Transaction Input, and Processing Location ID Code. Failure to do so results in error alert to system operations.

Measurement: Number of occurrences over specified periods of time. Confidence levels for attributes or sets cannot be known for sure.

Measurement Rule: Number of occurrences = number of application deficiency reports (DRs).

Assessment Rule: Existence of a DR.

Assessment: Red means the process is broken, Green means all systems are a go.

One MOIE objective was the use of the logical DQ metamodel as the basis for automatic generation of the schemas/sub schemas of physical database model. The features of the physical data model include the generation of the physical table specifications (i.e., creation of tables for entities); create relationships into foreign keys (i.e., needed for table relationships); convert attributes into columns, and de-normalize or modify the physical data model based on physical constraints/user requirements.

The DQ logical metamodel was created using data definitions extracted from ProVision v5.1.2. We attempted to migrate the logical metamodel from ProVision using the data definition language (DDL) migration interface to generate a physical data model. Multiple problems were encountered during this migration, one of which was the creation of foreign keys. In many database designs, one-to-many relationships are designated in the "many table" by naming the field the name of the other table, with a foreign key appended to the name. The ProVision interface generated this; however, it also generated some alien fields that appended to the foreign keys. More detail on the physical database built and usage has been addressed in a separate white paper on this subject.

2.3.3 Identification of Data Sources

To verify our postulates, we elected one area of specific interest to the AF and DoD, invoice transaction processing (See Figure 6). A recent General Accounting Office (GAO) report[12] cited numerous issues; some are business process related while others are attributed to DQ issues (e.g.,

[12] GAO Report: GAO-06-358, *DoD Payments to Small Businesses – Implementation and Effective Utilization of Electron Invoicing Could Further Reduce Late Payments*, May 2006.

timeliness and accuracy). The GAO noted, "ten percent of all invoices to large contractors" and "14.5 percent of small business invoices were paid late". Although the DoD has reported significant improvements in late payment metrics, these improvements have come through dedicating additional resources to the problem. They have yet to address the underlying root causes (i.e., DQ issues and processing weaknesses) that contributed to the late payments.

The DoD's response to the GAO findings resulted in the creation of a centralized, net-centric, Wide Area Work Workflow (WAWF) capability. WAWF has great potential of expediting invoice processing and has improved the overall timeliness of DoD payments to vendors. However, this is only a partial solution. Many other invoice inconsistencies and reconciliation activities still remain. Specifically, tracking and monitoring of contractor delivery status and compliance with the terms and conditions of AF contracts. In many cases, the DoD continues to process mostly paper payment documents. This approach is plagued with redundant data entry; misplaced documents, and higher than necessary transaction processing fees. Ultimately, payment delays, inaccurate contractor performance reporting, and untimely closure of contracts are the end results of invoice DQ issues.

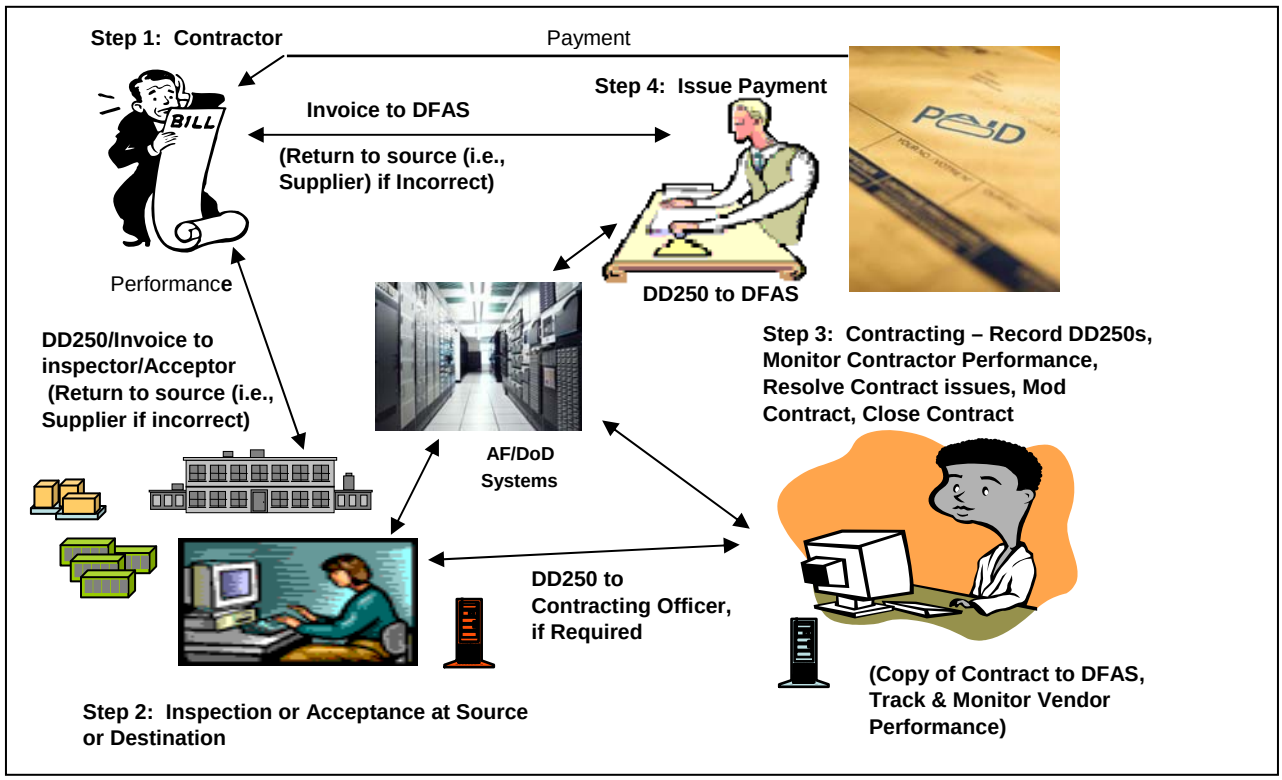


Figure 6. Invoice Transaction Flow

The MOIE team used the AF Acquisition and Due-in Asset Tracking System, ADIS-J041, as the source of AF invoice data. J041 is a mainframe, Common Business Oriented Language (COBOL) application implemented in September 1972. It provides tracking and reporting status on pre and post-contract award actions for Air Force Materiel Command (AFMC). AFMC has three ALCs that use the J041 system. The centers are located at Hill AFB, UT (OO-ALC); Tinker AFB, OK (OC-ALC); and Robins AFB, GA (WR-ALC). J041 processes over 120 plus different transaction formats. Additionally, this system processes and interfaces contract data (e.g., shipment status and asset visibility) to the Requirements Management, Stock Control, Maintenance, and Foreign Military Sales (FMS) customers/systems via the Contract Information Database System (CIDS - J018R). J041 processes invoice transactions daily providing item managers, procurement analysts, buyers, and case country managers with logistics and invoice information.

The J041, a batch processing system, processes incoming invoices from numerous sources (i.e., external interfaces or manual input). Those invoices that fail system edit and validation are flagged with error codes that identify the error conditions, which are reported to contract specialist /coding clerks for correction. As illustrated in Figure 7, the J041 system logic recycles and reprocesses these invoice transactions until they are either deleted or the rejection reasons are corrected

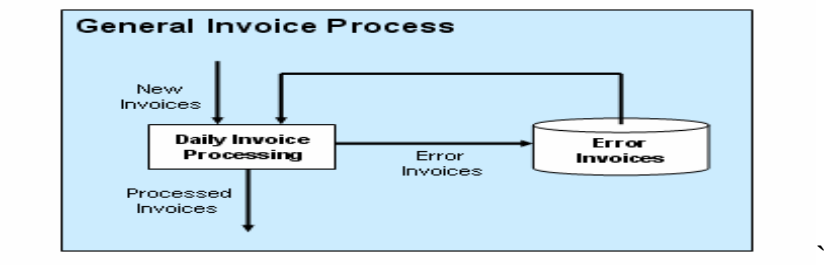


Figure 7. J041 Processing Flow

2.3.4 Data Capture

For this MOIE effort we elected to capture two months (9 Nov 2006 – 11 Jan 2007) of J041 invoice transactions. Figure 8 provides an overview of the high-level steps involved in the capture, extraction, analysis, loading and DQ reporting. A further elaboration on the activities performed to accomplish each step is defined in Figure 8.

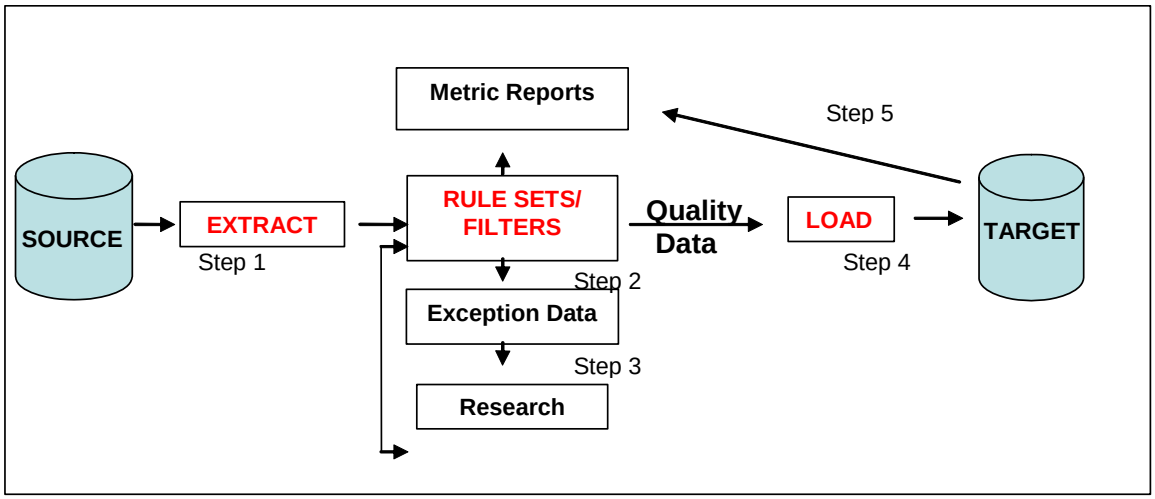


Figure 8. Invoice Extraction and Load Steps[13]

2.3.4.1 Step 1: Data Extraction

The J041 surveillance programmers provided copies of the valid and invalid invoice transaction files. Processing cycles included: daily, end of month (EOM), end of quarter (EOQ), and end of calendar year (EOCY) processing. The DQR for EOM, EOQ, and EOCY were not explored; however, it is important to note there may be business rules which provide for information exchanges with other systems on a periodic basis. The DQ dimensions coupled with the information exchange requirements between business applications can provide interesting insights into threads and impacts DQ throughout the AF.

In the absence of automated tools, valid and invalid invoices (i.e., PJJ/EJJ – transaction identifiers for invoice establish, update, or deletion actions) were initially captured in Microsoft Excel files. This approach was very cumbersome and labor intensive. However, it imperative this be done in a timely manner in order to educate the MOIE Team on the dynamics and complexities of AF invoice transaction processing.

2.3.4.2 Step 2: Profiling Data Exceptions

The profiling of invoice data included the capture of the data pedigree, invoice governance, identification of the transaction attributes (e.g., data element names, definitions, business rules, error conditions, descriptions) and the mapping to the DQ attributes. The business rules (See Appendix A) for the invoice transactions processing in J041 were obtained from system support personnel and supporting user and system documentation. This proved to be a very daunting task since the supporting system documentation was found to be out of date with the processing logic

[13] *DoD Guidelines on Data Quality Management*

contained within the J041 system. There is no controlled vocabulary registration authority for a government invoice (e.g., DD250 form). Thus, there is no unambiguous, non-redundant definition source to verify the data, metadata, business rules, and rationale for the J041 processing logic.

The MOIE Team evaluated automated data profiling tools from various vendors (e.g., IBM, SAS, Informatica, Business Objects, QBase) to assess their techniques for unlocking the mystery of source data content and structure. We experimented with the QBase's Exploratory Data Analysis EDA tool for data extraction, counting, and profiling (i.e., identification of occurrences of business rule violations) the invoice data. We discovered this tool significantly reduced the amount of time (compared to what was done in Step 1) for capturing counts (See Figure 9), error condition occurrences, in addition to providing a variety of different types of reports (e.g., Pareto, Standard Deviation and Variance, Probability of Distribution Kurtosis) across a population of observations. The Pareto Chart in Figure 9 shows the distribution by source of invoice transaction origin (i.e., "E" = external, "B" = manual input, and "L" = another ALC (i.e.,WR-ALC)).

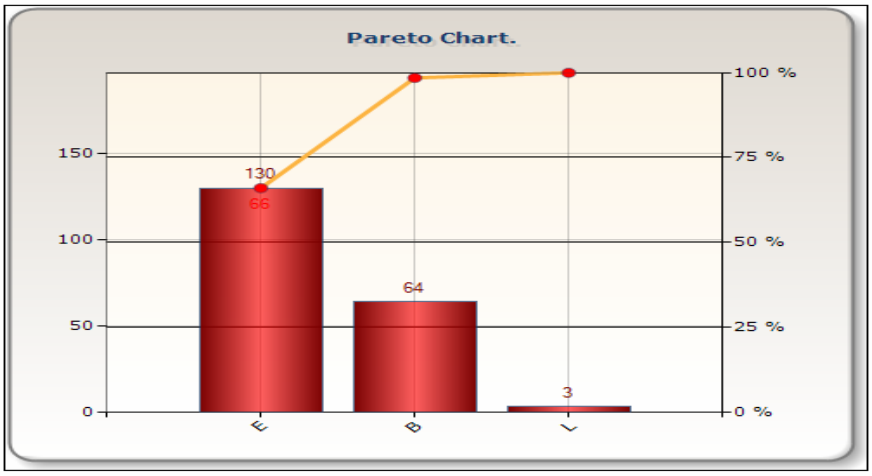


Figure 9. Invoice Distribution by Source

2.3.4.3 Step 3: Researching and Analyzing the Data

The invoice metadata was first used as the basis for a paper exercise on verifying and validating the metamodel research and design. The results suggested minor tweaks and changes, but begged a question about extensibility. Further research and analysis suggested that a DQ metamodel simply built around invoice metadata should be raised to a higher level of abstraction to be extensible to other types of data objects. Other observations included:

- Information can be frequently represented with different interpretations, satisfying differing user views and needs.

- We found that legacy systems are frequently a good place to gain an understanding of the real DQRs. It is important to understand that not all business rules are created equal. Therefore the real intent of a business rules is not always intuitively obvious. For example, a business rule applicable to the timeliness of invoice error correction can restrict the passing of invoice data to other systems or communities of interest, which in-turn activates business rules in other business applications based on the presence or absence of a shipment (e.g., invoice) actions. A registered invoice vocabulary would have been beneficial. Apparently there is an Open Applications Group Integration Specification (OAGIS) business object document for an industry standard invoice; however, this has not been embraced by the AF/DoD.
- Other important sources of DQRs include: Information Exchange Requirement or Interface Control Document specifications between systems; governance (e.g., policies, procedures, public law); and subject matter experts. It was discovered that context knowledge about data/information may be scattered across business domains because of data ownership/stewardship conflicts.
- When capturing the actual invoice DQ it was necessary to determine what other DQ attributes (e.g., age of the invoice, duration of time to correct the error, change of invoice transaction state for error to be corrected or erased/deleted) of interest. It is critical to establish the enterprise DQR, to understand the assessment rules and metrics which are used as the measurement baseline to determine the quality of the actual data. For invoices, the enterprise can be the DoD or the AF. It depends on who is the source of the requirement for goods and services and the organization responsible for issuing and administrating the contract and subsequent contract actions. When multiple organizations (e.g., government and commercial) are involved, there must be cross organization cooperation to ensure the accuracy, integrity, objectivity, and consistency of the data. Understanding these nuances are essential for capturing the rationale, business and processing rules, and DQ categories; all of which are needed for the identification and mapping of DQ dimensions for assessing the business impacts attributed to DQ issues.

2.3.4.4 Step 4: Populating the Metadata Repository (MDR) with Invoice Data

We utilized a SOA as the foundation for populating the MDR. A combination of services and publish/subscribe technologies, implemented using Mule (See Figure 10) as an enterprise service bus (ESB), are the backbone of the DQ prototype. A publication service allows data providers to specify the information products being delivered by the application, and indicate the relevant DQ metrics for this data. Once a publication is setup, the application publishes data to the system where it is processed and inserted into the MDR. The metadata repository management service receives the information product, generates data objects, and performs measurements for each metric defined for the data objects.

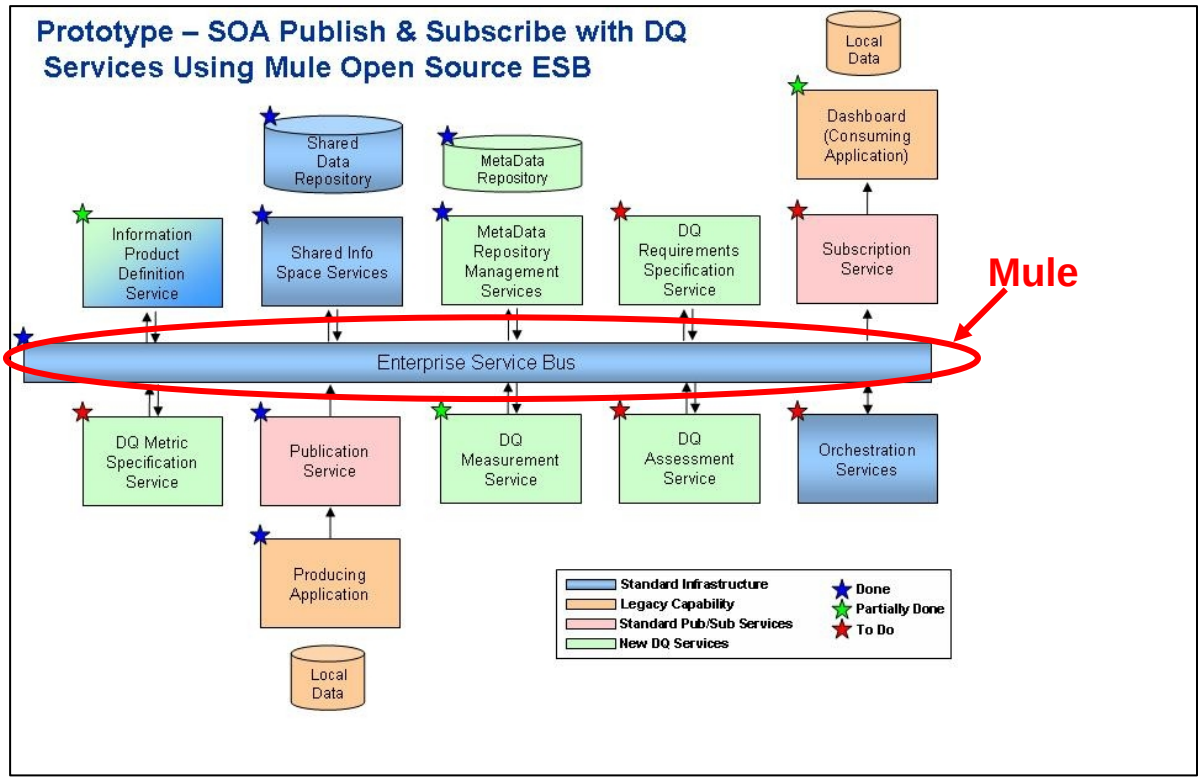


Figure 10. Mule Enterprise Service Bus

A subscription service allows the end user to discover information products available and subscribe to receive data from those publications. During subscription setup, the user also specifies DQR for the data object metrics. Before data is displayed to the user, a DQ assessment service compares the actual measurements to the requirements and generates the display using a red/yellow/green coloration scheme. A web-based client provides the user interface for publication and subscription setup and for displaying the assessment results for subscriptions. A separate white paper[14] exploits the underlying details of implementing an MDR in an SOA environment.

2.3.4.5 Step 5: Metrics and Measurements

Assuming the role of a data quality manager in Contracting, we considered a strategy for improving the quality of invoice data. We made a determination of what distinguished good quality data from bad quality data in measurable terms. Answering this question helped to identify and describe the DQ requirements. Once we understood the DQ requirements and the aspect of DQ that we wanted to measure, we created measurement rule(s) and assessment rule(s)

[14] *Implementing Data Quality in a Publish and Subscribe Service Oriented Environment*, by Michelle Casagni, Kelly Whitacre, Geoff Parsons, Jacob Fenwick, and Dave Becker.

along with the assessment scoring criteria (i.e., below 20 = red, 20-30 = yellow, 30-40 = green, above 40 = blue, etc.). All of this metadata gave us the heuristics that appeared most appropriate for incorporation in the DQ logical metamodel. This provided us with the basis to access AF invoice DQ.

2.3.5 Identify DQ reporting requirements and dash board DQ reports

While creating the conceptual and logical DQ metamodels, we created some notional use cases perceived to be of value to the DQM. We assumed the need for metrics that would aide in reducing the longevity of invoice processing cycles and for the reduction of costs associated with invoice processing (e.g., labor, system, no common invoice media [i.e., mail, fax, automated transactions in various formats, cost of money, etc.])[15]. We studied the cause and effects of the various J041 invoice exception reasons. The following is a cause and effect list of invoice issues which were considered as candidates for use case scenarios used to validate the DQ metamodel.

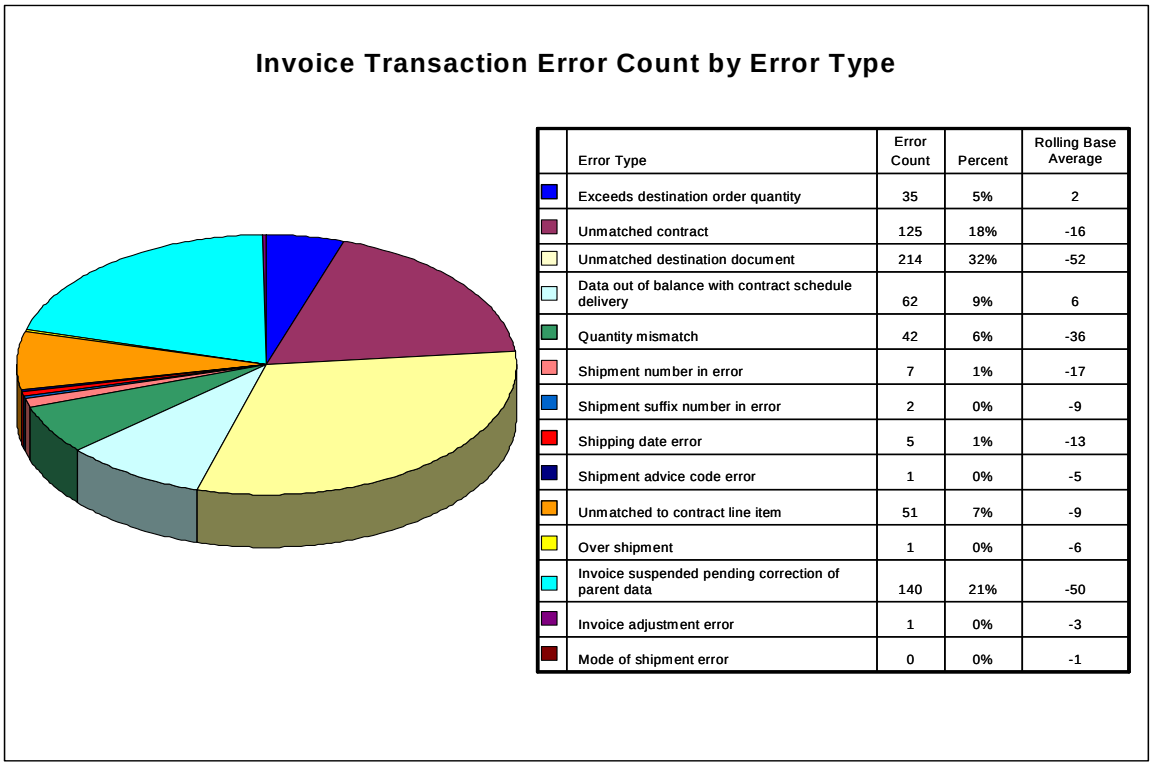
1. Invoice transaction errors caused by the vendor:
    a. Delays in government Inspection and Acceptance process can lead to delays in delivery
    b. Frequent billing and payment errors (e.g., duplicate payments)
    c. Assets to be shipped to the wrong destination (e.g., value of assets lost in route, cost of product returns or redistribution, and creates delays in delivery)
    d. Deliverables are not IAW contract terms and conditions (e.g., wrong items or color, packaging, inaccurate fit, form, function, and incorrect quantity shipped) which leads to contractor performance and creates delay in delivery issues
    e. Created the loss of time/resources expended researching/resolving records of discrepancy
    f. Untimely processing of invoices which may increase awaiting parts status and reduce weapon system availability rates
    g. Generate the need for additional assets in pipeline to compensate for supplier shortfalls

[15] Source: DFAS/MOCAS rates for 2006 per contract line item: manual processing = $33.98/automated transaction processing = $20.42. Commercial rates per invoice: manual processing = $25.32/automated transaction processing = $3.97. According to an IOMA study conducted in 2003 (A/P Department Benchmarks and Analysis), the average accounts payable department takes 6.8 days to process an invoice.

2. Untimely processing of invoices caused by the government:
    a. Results in late payment penalties (e.g., interest payments) and lost opportunities to capitalize on invoice discounts
    b. Causes legacy systems to erroneously report vendor performance data, erroneous tracking/reporting of asset due-ins, and delays in contract closeout
    c. Increases labor burden, manual input of invoices is prone to human error. If not immediately rectified, additional time is expended in back tracking and researching rationale for discrepancies[16].
    d. Creates inconsistency in invoice processing, tracking, and reporting.

Figure 11 provides a graphic of displaying the consolidated error counts for a given day by error type. It shows whether there is a decrease or increase from a rolling baseline average. A cursor over a pie slice provides the user with a breakout count by ALC and by input source. This provides the DQM with a "quick look" at data quality trends. Graphical reports are recommended to provide a comparative basis of data quality.

[16] Source: Unknown, a frequently used number in ALC circles is that on average, an invoice costs $15-$50 to process.



Error Type	Error Count	Percent	Rolling Base Average
Exceeds destination order quantity	35	5%	2
Unmatched contract	125	18%	-16
Unmatched destination document	214	32%	-52
Data out of balance with contract schedule delivery	62	9%	6
Quantity mismatch	42	6%	-36
Shipment number in error	7	1%	-17
Shipment suffix number in error	2	0%	-9
Shipping date error	5	1%	-13
Shipment advice code error	1	0%	-5
Unmatched to contract line item	51	7%	-9
Over shipment	1	0%	-6
Invoice suspended pending correction of parent data	140	21%	-50
Invoice adjustment error	1	0%	-3
Mode of shipment error	0	0%	-1

Figure 11. Invoice Transaction Error Count by Error Type

The graphic in Figure 12 is an aggregation on invoice errors over a period of time by location, using red, yellow, and green assessment criteria. This report shows how many standard deviations a specific location is from the mean. For example, of the 2475 invoices processed at WR-ALC between 9 Nov 06 and 11 Jan 07, only 18% of them were rejected, which is below the mean, thus a green rating. The OC-ALC processed 3962 invoices of which 44% (1737) of them were rejected in error. This is above the mean and one SD, thus a red rating. From an AFMC Major Command perspective, 32% (2542 out of 7935) of invoices were rejected. This would suggest DQM needs to investigate the reasons why one location appears to have a higher invoice processing success rate compared to the other two locations.

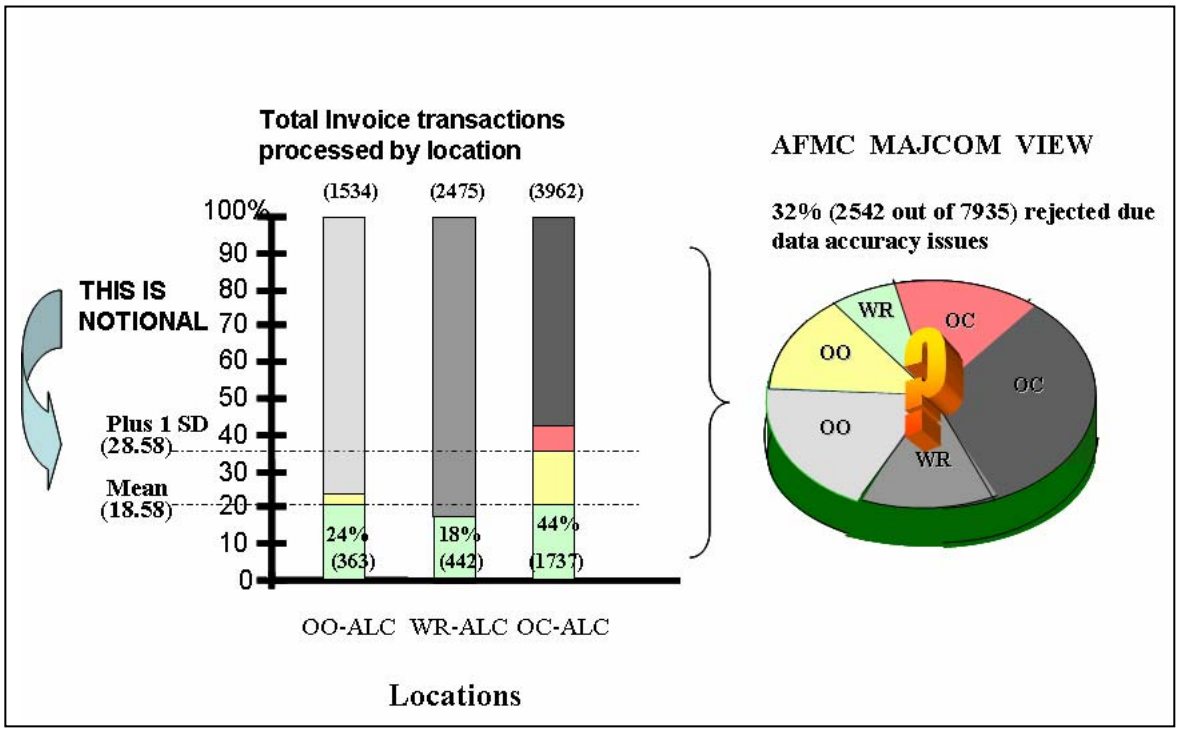


Figure 12. Percentage of Invoice Errors by Processing Location for 9 November 2006 – 11 January 2007

2.3.6 Design and build a DQ infrastructure

The design and build of a DQ infrastructure was researched and accomplished by other members of the MOIE Team. A separate white paper is being developed to address challenges, a solution, and lessons learned from creating to DQ SOA environment.

2.3.7 Demonstrate a systematic approach

The demonstration of the DQ SOA environment is a work-in-progress as of the published date of this report. A separate white paper is being developed, documenting the viability of the SOA environment and providing examples of dashboard reports on the DQ of the AF invoice data.

2.4 Observations, Findings, and Recommendations

2.4.1 Other DQ issues discovered while profiling and analyzing J041 data

While profiling and analyzing the J041 valid and invalid transaction files, a number of other questionable transactions and transaction behaviors, perceived to be data quality issues, were discovered. Examples include the following:

- Garbled or special characters interpreted to be garbage transactions: This appeared to be a common occurrence (ref BZI56b file 9 Nov 06 contained "/////" in the 80 character transaction. The transaction(s), created on day 313 of 2006, rejected correctly with a "1AB" exception reason.

- Inconsistent transaction formats: Invoice formats are being reformatted (OC-ALC only). Instances of PJJ transactions processing with exception reasons are being reformatted and processed with a "GEY" Routing ID assigned. An undefined data element appears in columns 200-203, content values vary, e.g., "7CHC", "JAEC". This occurs daily, reference bzi56b file dated 14 Nov 06, Contract number F3460103D00410021, Line Item 0009AD.

- Internally inconsistent data*:* OO-ALC Revised Delivery Forecast "PJA" transactions (from 15 Nov 06 processing cycle) have a special character (e.g., "ÿ" in the primary key of the system generated sort key which is appended to the 80 column input transaction. The origin of the key data comes from the 1-80 column input which suggests there is a system sort key data build error.

  Example: "PJAFA821206C0012PM0043U00P U 08MAY31……GDDEL, sort key = FA821206C0012 20U00P ÿU ……. 2106312B20 05 7PL 08152……." This condition is consistent across the three ALCs, however inconsistent with system support documentation.

- Abnormal processing intervention: Manipulation of data outside the normal system manufacturing/processing flow. For example: Unusual invoice transaction processing behavior occurred at OC-ALC on 5 Dec 06 (reference Appendix B – Daily Invoice Processing Counts). There were 631 invoice error transactions suspended in J041 at the end of the 4 Dec 06 processing cycle. On 5 Dec 06, they all disappeared without cause (i.e., no invoice deletion/erase transactions processed). No explanation was given why these invoices disappeared, nor an explanation for why 566 on the 631 reappeared on 7 Dec 06, and another 14 of the original 361 reappearing on 11 Dec 06.

- Obsolete data: Data has been overcome by events and the data has not been updated or deleted in a timely manner. For example: On 5 Dec 06, at OO-ALC, over 6300

Revised Delivery Forecast (RDF) transactions are recycling in the J041, input from the Provisioning System, D220. The bulk of these transactions where for contract line items (appeared to be for line items with "u" unknown delivery schedule dates and transactions are establishing delivery dates that are over three years old) are no longer in J041. Example: "PJAF4262099C0027PM0198H010AA U 03JUL31 GDDELF4262099C0027 20H010AAÿU 2102220B202FC 93 7PL 03212 7" The RDF date is "03JUL31"

The above examples pose a number of interesting DQ observations. The timeliness to which WR-ALC researches and corrects invoice error transactions (See Figure 12) is far superior to the other two ALCs. This trend appears to be consistent for all other J041 transaction error conditions. OO-ALC transaction error counts, over a two-month timeframe, varied from 18K – 25K daily, the majority of which are recycled from one processing cycle to the next. The OC-ALC counted was in the 1200-5K range. Random observations of the age of OO-ALC error transactions suggest 65% are greater than one year old. OC-ALC J041 error transactions are in the 30-35% one year or older range, WR-ALC is around 20%, mostly attributed to the "MRT/MWT" (miscellaneous items) errors.

The more important question, what is the overall impact on the AF and its strategic partners who depend on the timeliness, completeness, precision, consistency, and accuracy of J041 data? The J041 system, via the J018R Contract Information Database System (CIDS), provides contractor asset delivery (i.e., due-in asset visibility) data to other AF users (e.g., item managers and buyers) and a large number of other information systems (e.g., inventory management, logistics, FMS, requirement computation, and business intelligence). These systems compute information based on the presence or absence of J041 data. The systems generate reports to their users, and share J041 data with a host of other systems. All of these various organizations and their support system contribute to the data manufacturing process of the AF and other Service entities.

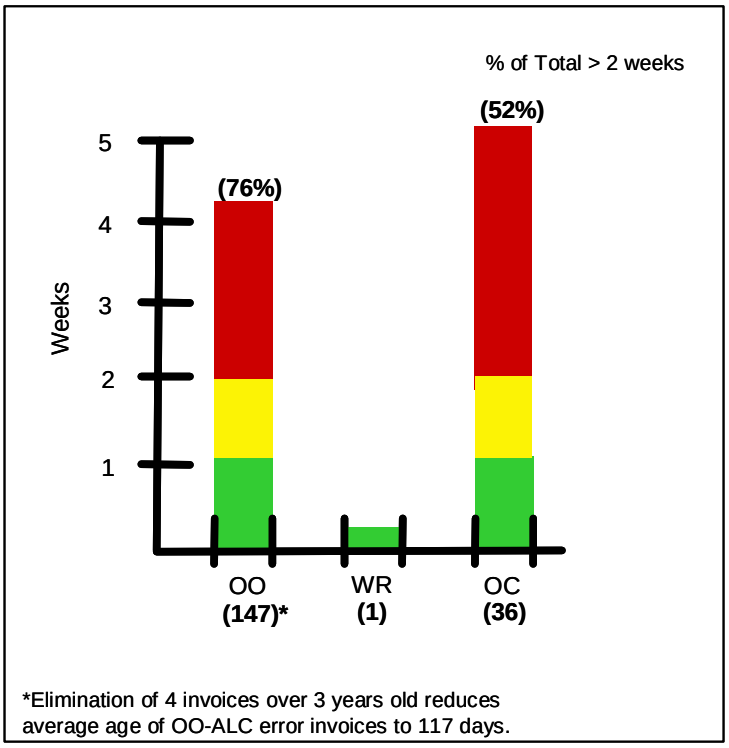


Figure 13. Average Age (days) to Correct Invoice Errors

The efficiencies and effectiveness of managing the AF business activities are largely dependent upon the dimensions of DQ. The examples in Figure 13 are just a few of the factors that drive the costs of AF operations (e.g., number of assets in the pipeline, low awaiting parts, and high aircraft availability rates). Being able to quantify, qualify, and measure DQ is critical to understanding the overall impacts to the AF mission, that of sustaining and supporting the warfighter. The results from this MOIE suggest the AF needs to place more emphasis on DQ.

2.4.2 Recommendations

The AF may want to consider establishing quality guidelines that reflect the true value stream of timely invoice processing to the AF enterprise. MITRE recommends the following quality goals, objectives, and measurements:

- Reduction of invoice-processing errors and costs – the cost of manual labor drives up the price for invoice transaction processing. The key is finding the best of breed/right balance of technology, process, and trained workforce that will result in the timely delivery of quality invoice information at a lower cost. The goal would be to strive for a six-sigma level of consistency where there are only 4.3 defects per million invoices (Measure = man-hours and dollars expended to resolve invoice errors). The

return on investment is in the benefits of processing invoices in a timely manner and freeing up resources to expedite conflict resolution on any other transaction issues.

- Maximize emphasis on obtaining early payment discounts – earning discounts for early payment is a definite way to minimize costs. This requires a capability to quickly identify invoices and extracts of the required information needed so the Defense Finance and Accounting Service (DFAS)/Payment Office can issue payment and gain the discount. The goal is to realize the discounts 99% of the time they are offered on government contracts. Measure = number of instances (i.e., contract line items where invoices are processed in a timely manner) where prompt payments are made at the discounted rates.
- Data Accuracy – Acquiring the mechanisms (i.e., invoice workflows) and establish the incentives to obtain 99% plus accuracy is a critical first step in increasing the efficiencies (i.e., increases productivity of government resources) and effectiveness of invoice processing. Measure = determine the total number of invoice transactions processed with data errors compared to the total population of invoice transactions.
- Improved vendor relationships and customer service – Just-in-time delivery of quality goods and services and prompt payments occurring 99% (arbitrary %) of the time achieves improve effectiveness and satisfaction of all parties (supplier and customer). The goal is to reduce the costs associated with awaiting parts and customer wait times. Measure = sales order performance compared to customer's due date.

When practical, utilize commercial best practices (e.g., Wal-Mart invoice verification/validation and/or Federal Express package handling, transportation, and locator models). The terms and conditions of the ALC contracts should require vendors/contractors to utilize Automatic Identification Technology (e.g., OCR and ICR recognition technology for highly accurate invoice data capture, including: vendor ID, date, contract line item, shipment number, quantity shipped, and order item attributes). This information can be captured in real-time and automatically delivered to the AF for validation and verification against the terms and conditions of the contract. The goal should be to identify and correct invoice errors as possible prior to the material leaving the vendors/suppliers' facilities. Vendor performance should be tied to invoice accuracy and timeliness.

All vendor invoices should be cleared through a central clearinghouse (e.g., the WAWF concept – one true invoice source) requiring compliance through a strictly audited, workflow process. Every invoice should have a fingerprint. This would invoke a common set of invoice edit and validation rules versus having numerous legacy systems, each doing their own thing, perpetuating point-to-point information exchanges to other systems, thus promoting other transaction/actions which may or may not be valid. Leverage the use of technologies that capture a variety of document types, including multi-lingual documents. The goal would be to implement a quality-at-the-source

practice with vendors and government sources and leverage the use of technology to minimize human intervention in the processing of invoices.

The recommendations below are based on achieving the following quality goals, objectives, and suggested measurements:

- Reduced invoice-processing errors and costs – The cost of manual labor drives up the price for invoice transaction processing. The key is finding the best of breed/right balance of technology, process, and trained workforce that will result in delivering quality invoice information (e.g., error reduction) at a lower cost. The goal would be to strive for a six-sigma level of consistency where there are only 4.3 defects per million invoices (Measure = man-hours and dollars expended to resolve invoice errors). .
- Maximize emphasis on obtaining early payment discounts – Earning discounts for early payment is a definite way to minimize costs. This requires a capability to quickly identify invoices and extracts of the required information needed so the DFAS/Payment Office can issue payment IAW the discount terms and conditions in AF contracts. The goal is to realize the greater utilization and benefits from discounts provisions contained in AF contracts. Measure = number of instances (e.g., contract line items where invoices are processed in a timely manner) where prompt payments are made at the discounted rates.
- Data Accuracy – Acquiring the mechanisms (i.e., invoice workflows) and establishing the incentives to obtain 99% plus accuracy is a critical first step in increasing the efficiencies (i.e., increases productivity of government resources) and effectiveness (i.e., improves timeliness) of invoice processing. Measure = determine the total number of invoice transactions processed with data errors compared to the total population of invoice transactions.
- Improved vendor relationships and customer service – Just-in-time delivery of quality goods and services and prompt payments occurring 99% (arbitrary %) of the time achieves improved effectiveness and satisfaction of all parties (supplier and customer). The goal is to reduce the costs associated with awaiting parts and customer wait times. Measure = Sales order performance compared to customer's due date.

3 Conclusions

Although MITRE's research only used the J041 invoice data, the DQ findings of other DQ assessments[17] suggest DQ issues are prevalent in AF legacy data systems. The threads of effects

[17] *Data Quality Processes: Insurance Against Negative Operational Impacts;* Presentation given by Fred Nannarone, Dynamics Research Corp, ECSS Data Summit, August 2007

of DQ issues are not intuitively obvious and next to impossible to trace (i.e., lack of ancestral relationships) to the root cause(s). Frequently, the effects of DQ issues are hidden because of user manipulation (i.e., manual and automated work-arounds). The continued masking of these DQ issues may suggest the immutable law of nature, which states "garbage in garbage out", has been repealed. We know this is not the case. Sooner or later the prevailing techniques and methods of addressing DQ issues will have to change. Pursuing a pro-active course of action of applying enablers (i.e., people with good analytical skills and tools) can identify, categorize, trace, and quantify the degree of data trustworthiness in the AF legacy systems.

The lead time necessary to discover, analyze, and institute corrective action is rapidly evaporating. Critical manpower resources are soon to retire and the window of opportunity is closing to correct DQ issues. These enviable events will lead to the temptation to migrate legacy system data. Lessons learned from other DoD agencies (e.g., lack of DQ awareness in the Army caused a delay in their ERP deployment)[18] and commercial ERP deployments echo the essentials of having good data quality. Zero data defects are no longer just a program, but an important practice that must be enforced. A strategy of migrating legacy system dynamic (i.e., transaction) data as a last resort may be the best approach for the AF. All legacy system static data should only be considered after careful evaluation of the DQ requirements and characteristics in the context in which MITRE has modeled in this MOIE research.

MITRE recommends the AF accelerate the adoption of a disciplined DQ approach. Functional data stewards should step up to the responsibility of analyzing and documenting the changes (i.e., roles, business rules, impacts on process behavior as it pertains to data) of business process re-engineering efforts and educate legacy system sustainment personnel on what data will be needed for future solutions. It is very important to start fostering enterprise-wide mind-sets and focus on the core data needed to support the warfighter. The benefits from having a DQ strategy will help clean up the data, improve operations control, improve business intelligence, analytics, and decision support, facilitate business process re-engineering, and six-sigma initiatives.

[18] Mr. Michael Gallagher, retired Army CIO, take away from ECSS Data Summit, Aug 2007

Appendix A Invoice Data Quality Requirements (i.e., J041 Error Conditions)

Error Code	Error Description	Business Rule
A73	INVOICE MODE OF SHIPMENT	Mode of shipment code must be A-Z, 2-9 *, or %
1AJ or A76	SHP/RCPT DATE	Must meet date table requirements (See A03). When D4 receipt has been received from D035/DO33/DO34A containing a Julian date which is not convertible to an in-the-clear date. The invalid date has been placed in the first four, positions (52-55) of the PRN. NOTE: This exception will be assigned an A76 exception code the next J041 processing cycle.
A77	SHIP/RCVD QUANT	Must be filled numeric and greater than zero
A78	SHIPMENT SUFFIX	1. Positions 4-9 must be FD2020, FD2030, FD2040 FD2050, or FD2060. 2. Positions 10-11 must be numeric. 3. Positions 12-16 must be filled. 4. Positions 17-22 must be blank. 5. Line item - positions 23-26 must be filled. If positions 27-28 are entered, both must be alpha or 27 may be alpha and 28 is blank.
2FA	UNMATCHED CONTRACT	Transaction does not match an existing contract master.
2FC	UNMATCHED CONTRACT	Transaction does not match an existing contract line LINE ITEM item master.
2FE	INVALID ADJUSTMENT	1. PWJ/PWN is unmatched to an existing shipment master. 2. PJJ either shipment advice code C or D is unmatched to an existing shipment master. 3. PWN or PJJ (advice code D) quantity exceeds the quantity on the matching shipment master.
2FJ	OVER SHIPPED	Shipment transaction over-ships destination order, quantity or line item quantity

Error Code	Error Description	Business Rule
2EV	NOT WITHIN UNDER RUN %	PJJ entered with a shipment advice code Z to indicate final shipment under-run and quantity shipped is not within the allowable under-run variance at line item level. This usually indicates a missing shipment.
2EG	UNMATCHED	Shipment/Receipt is unmatched to destination.
2eQ	PURCH UNIT MSTR	Shipment, Receipt purchase unit is different than the line item master purchase unit. Review and correct purchase unit. Note: quantity adjustment may be necessary.
2WC	SCHEDULE BALANCE	Line item established or changes to existing master would cause the schedule quantities and contract line item quantities to be out of balance.
PDE*	PARENT DATA ERROR	Invoice business rule not invoked until parent data errors are corrected.
* This error code was fabricated because J041 processing rules does not edit and validate invoice transactions (no invoice error code assigned for this condition) when the parent data is in error. This error code was created to ensure that all invoice transactions in error had a unique identifier for this condition.		

Appendix B Daily Invoice Processing Statistics

Date	OO-ALC								WR-ALC								OC-ALC							
	Tot Valid Inv Proc	Inv Input & Proc Today	Error Inv Input Today	Inv Errors Prev Days	Inv Input Today		Tot Population		Tot Valid Inv Proc	Inv Input & Proc Today	Error Inv Input Today	Inv Errors Prev Days	Inv Input Today		Tot Population		Tot Valid Inv Proc	Inv Input & Proc Today	Error Inv Input Today	Inv Errors Prev Days	Inv Input Today		Tot Population	
					Tot % Proc Valid	Tot % Proc Error	% Proc Valid	% Proc Error					Tot % Proc Valid	Tot % Proc Error	% Proc Valid	% Proc Error					Tot % Proc Valid	Tot % Proc Error	% Proc Valid	% Proc Error
9-Nov-06	10	10	3	191	77%	23%	5%	95%	51	51	17	0	75%	25%	75%	25%	24	24	103	331	19%	81%	5%	95%
13-Nov-06	16	16	4	193	80%	20%	8%	92%	40	31	9	0	78%	23%	82%	18%	247	247	9	345	96%	4%	41%	59%
14-Nov-06	22	22	11	188	67%	33%	10%	90%	34	33	10	1	77%	23%	76%	24%	90	85	11	347	89%	11%	20%	80%
15-Nov-06	13	13	11	191	54%	46%	6%	94%	91	91	23	0	80%	20%	80%	20%	36	35	42	349	45%	55%	8%	92%
16-Nov-06	25	24	3	199	89%	11%	11%	89%	61	53	9	0	85%	15%	87%	13%	59	59	22	312	73%	27%	15%	85%
17-Nov-06	44	44	8	198	85%	15%	18%	82%	31	31	9	0	78%	23%	78%	23%	35	34	27	389	56%	44%	8%	92%
20-Nov-06	26	26	6	198	81%	19%	11%	89%	40	40	2	0	95%	5%	95%	5%	39	38	34	374	53%	47%	9%	91%
21-Nov-06	10	10	3	200	77%	23%	5%	95%	126	126	16	0	89%	11%	89%	11%	51	51	58	395	47%	53%	10%	90%
22-Nov-06	22	22	22	203	50%	50%	9%	91%	123	115	21	0	85%	15%	85%	15%	36	36	58	407	38%	62%	7%	93%
25-Nov-06	0	0	4	215	0%	100%	0%	100%	0	0	0	16	0%	0%	0%	100%	0	0	22	414	0%	100%	0%	100%
27-Nov-06	42	42	16	198	72%	28%	16%	84%	41	35	10	0	78%	22%	80%	20%	87	83	51	424	62%	38%	15%	85%
28-Nov-06	34	34	12	205	74%	26%	14%	86%	3	3	4	2	43%	57%	33%	67%	65	65	99	426	40%	60%	11%	89%
29-Nov-06	40	40	8	213	83%	17%	15%	85%	18	18	1	2	95%	5%	86%	14%	41	41	85	445	33%	67%	7%	93%
30-Nov-06	6	6	2	216	75%	25%	3%	97%	13	13	0	2	100%	0%	87%	13%	0	0	30	492	0%	100%	0%	100%
1-Dec-06	61	61	7	193	90%	10%	23%	77%	133	133	17	0	89%	11%	89%	11%	178	178	143	509	55%	45%	21%	79%
4-Dec-06	0	0	3	198	0%	100%	0%	100%	121	117	9	3	93%	7%	91%	9%	101	101	69	631	59%	41%	13%	87%
5-Dec-06	66	66	25	195	73%	27%	23%	77%	51	45	3	0	94%	6%	94%	6%	44	44	3	0	94%	6%	94%	6%
6-Dec-06	56	56	24	208	70%	30%	19%	81%	6	3	3	3	50%	50%	50%	50%	34	34	3	3	92%	8%	85%	15%
7-Dec-06	13	13	1	219	93%	7%	6%	94%	28	24	24	4	50%	50%	50%	50%	30	29	43	566	40%	60%	5%	95%
8-Dec-06	30	30	14	219	68%	32%	11%	89%	18	10	18	8	36%	64%	41%	59%	0	0	0	0	0%	0%	0%	0%
11-Dec-06	27	27	6	227	82%	18%	10%	90%	82	78	8	0	91%	9%	91%	9%	52	50	32	580	61%	39%	8%	92%
12-Dec-06	30	30	5	207	86%	14%	12%	88%	44	42	3	0	93%	7%	94%	6%	23	23	41	584	36%	64%	4%	96%
13-Dec-06	22	22	10	201	69%	31%	9%	91%	108	106	11	0	91%	9%	91%	9%	59	52	40	581	57%	43%	9%	91%
14-Dec-06	24	24	10	202	71%	29%	10%	90%	116	113	16	1	88%	12%	87%	13%	48	48	37	373	56%	44%	10%	90%
15-Dec-06	14	14	10	209	58%	42%	6%	94%	49	46	5	1	90%	10%	89%	11%	53	53	51	350	51%	49%	12%	88%
18-Dec-06	48	47	4	199	92%	8%	19%	81%	11	11	3	1	79%	21%	73%	27%	21	21	76	356	22%	78%	5%	95%
19-Dec-06	34	34	1	200	97%	3%	14%	86%	180	180	43	3	81%	19%	80%	20%	52	52	59	358	47%	53%	11%	89%
20-Dec-06	44	43	14	196	75%	25%	17%	83%	60	50	38	21	57%	43%	50%	50%	83	83	36	366	70%	30%	17%	83%
21-Dec-06	66	66	7	201	90%	10%	24%	76%	33	33	5	14	87%	13%	63%	37%	85	84	56	375	60%	40%	16%	84%
22-Dec-06	19	19	7	208	73%	27%	8%	92%	14	14	13	16	52%	48%	33%	67%	73	71	55	376	56%	44%	14%	86%
26-Dec-06	24	23	17	211	58%	43%	10%	90%	29	29	9	22	76%	24%	48%	52%	38	38	46	387	45%	55%	8%	92%
27-Dec-06	6	6	2	212	75%	25%	3%	97%	23	18	3	19	86%	14%	51%	49%	115	58	18	258	76%	24%	29%	71%
28-Dec-06	29	29	8	212	78%	22%	12%	88%	90	90	21	21	81%	19%	68%	32%	75	75	8	264	90%	10%	22%	78%
29-Dec-06	0	0	1	220	0%	100%	0%	100%	11	8	7	25	53%	47%	26%	74%	0	0	7	5	0%	100%	0%	100%
3-Jan-07	47	47	6	218	89%	11%	17%	83%	3	3	0	30	100%	0%	9%	91%	0	0	0	0	0%	0%	0%	0%
4-Jan-07	44	42	21	220	67%	33%	15%	85%	23	23	3	30	88%	12%	41%	59%	64	64	28	5	70%	30%	66%	34%
5-Jan-07	46	44	9	235	83%	17%	16%	84%	27	27	3	33	90%	10%	43%	57%	34	34	37	5	48%	52%	45%	55%
8-Jan-07	48	45	5	231	90%	10%	17%	83%	10	10	13	34	43%	57%	18%	82%	30	30	48	31	38%	62%	28%	72%
9-Jan-07	17	9	2	220	82%	18%	7%	93%	49	44	20	20	69%	31%	55%	45%	15	15	82	45	15%	85%	11%	89%
10-Jan-07	34	34	17	191	67%	33%	14%	86%	118	109	13	1	89%	11%	89%	11%	63	63	48	5	57%	43%	54%	46%
11-Jan-07	32	31	14	195	69%	31%	13%	87%	27	27	0	12	100%	0%	69%	31%	91	91	20	7	82%	18%	77%	23%

Appendix C Glossary

AF	Air Force
AFMC	Air Force Material Command
ALC	Air Logistic Center
CIDS	Contract Information Database System
COBOL	Common Business Oriented Language
DDL	Data Definition Language
DFAS	Defense Finance and Accounting Service
DoD	Department of Defense
DQ	Data Quality
DQA	Data Quality Architecture
DQM	Data Query Management
DQR	Data Quality Requirements
DR	Deficiency Reports
EAI	Enterprise Application Integration
EOCY	End of Calendar Year
EOM	End of Month
EOQ	End of Quarter
ERP	Enterprise Resource Planning
ETL	Extract and Load
FMS	Foreign Military Sales
GAO	General Accounting Office
IAW	In accordance with
IER	Information Exchange Requirement
IT	Information Technology
MDR	Metadata Repository
MOIE	Mission Oriented Investigation and Experimentation
OAGIS	Open Applications Group Integration Specification
OC-ALC	Tinker AFB, OK Air Logistic Center
OMB	Office of Management and Budget
OO-ALC	Hill AFB, UT Air Logistic Center
OWL	Ontology Web Language
PJJ/EJJ	Transaction identifiers
RDF	Resource Description Framework
RDF	Revised Delivery Forecast
SD	Standard Deviation
SOA	Service Oriented Architecture
WAWF	Wide Area Working Flow

WR-ALC	Robins, GA Air Logistic Center
XML	eXtensible Markup Language